Scaling the Earth System Grid to 100Gbps Networks

End-to-End Data Delivery and Management in Extreme Scale

Alex Sim, Mehmet Balman, CRD, LBNL
Dean N. Williams, PCMDI, LLNL

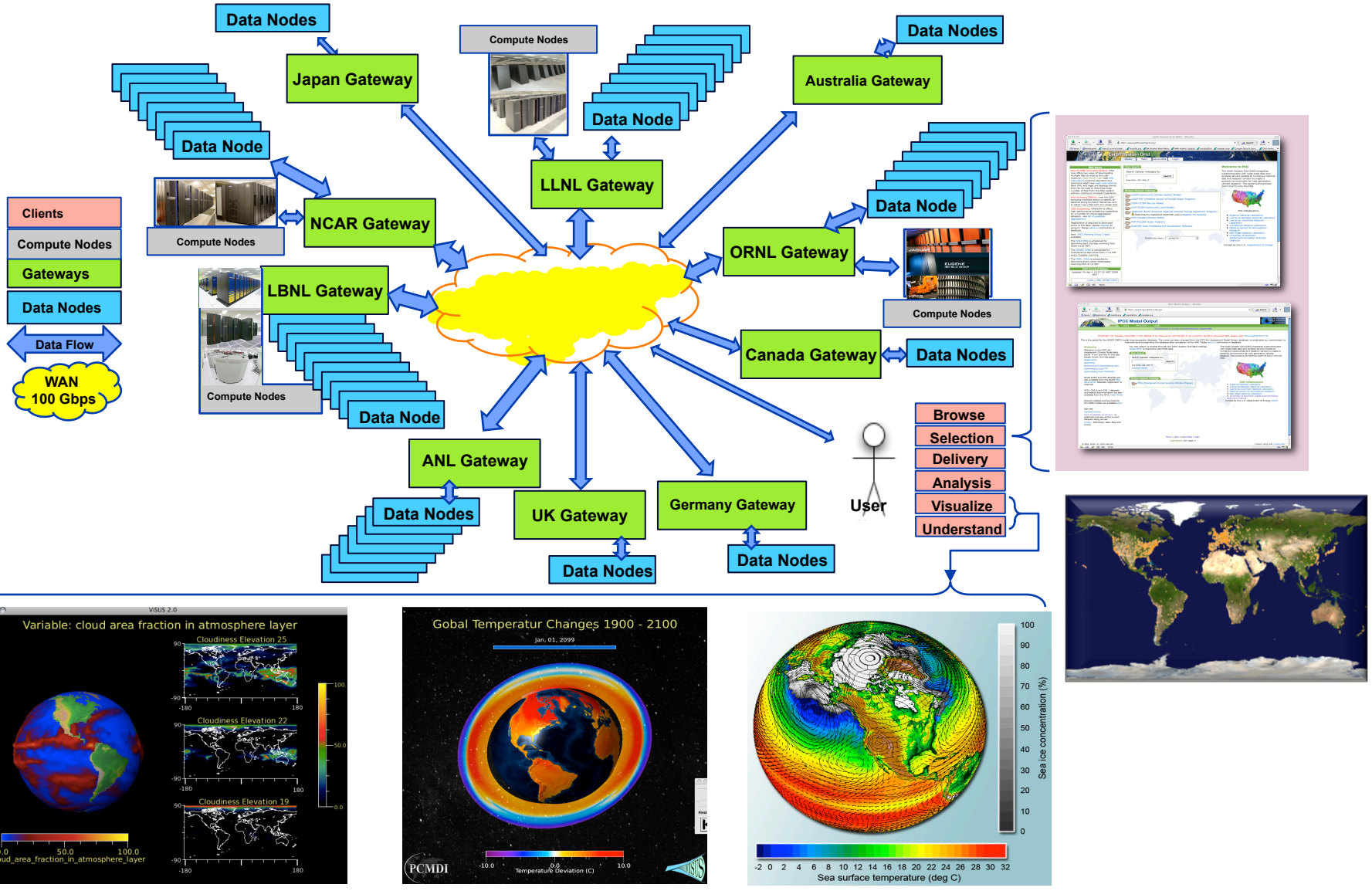
Overview of project

- **ANI Research project**
  - **Started 9/2009**
    - **Project ends in 9/2012**
  - **Supports one postdoc at LBNL**
  - **Study high performance network scalability in Earth System Grid for distributed data access and replication**

100Gbps Network and Climate Community

- **High performance network for climate community**
  - **Distributed data access and replication around the world**
    - **Meet the needs of national and international climate projects for distributed datasets, data access, and data movement.**
    - **Integrate highly publicized climate data sets using distributed storage management, high-performance analysis environment, and high-bandwidth wide-area networks.**
    - **"Replica Core Archive" – The Coupled Model Intercomparison Project, Phase 5 (CMIP5) used for the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5) is estimated to 1.5-2PB.**
    - **Climate model data is projected to exceed hundreds of Exabytes by 2020**
- **Climate100**
  - **Research effort for 100Gbps network usage from data intensive applications point of view**
    - **Enable new capabilities for analysis of data and visual exploration**

Scaling climate community to 100 Gbps networks

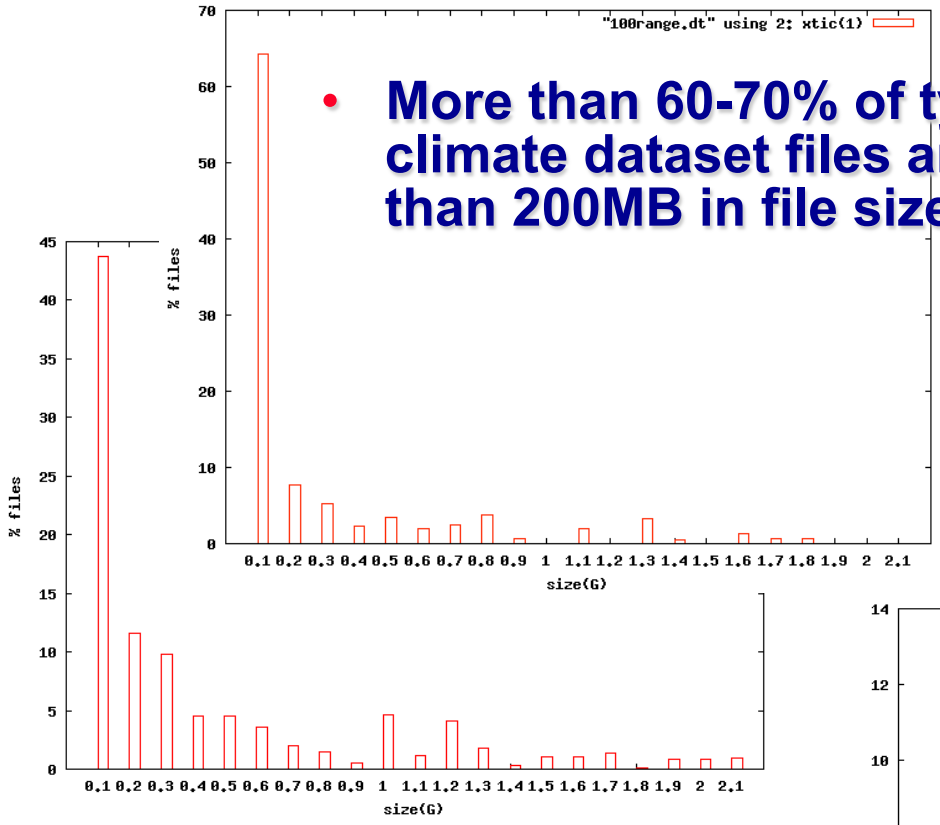


Efforts

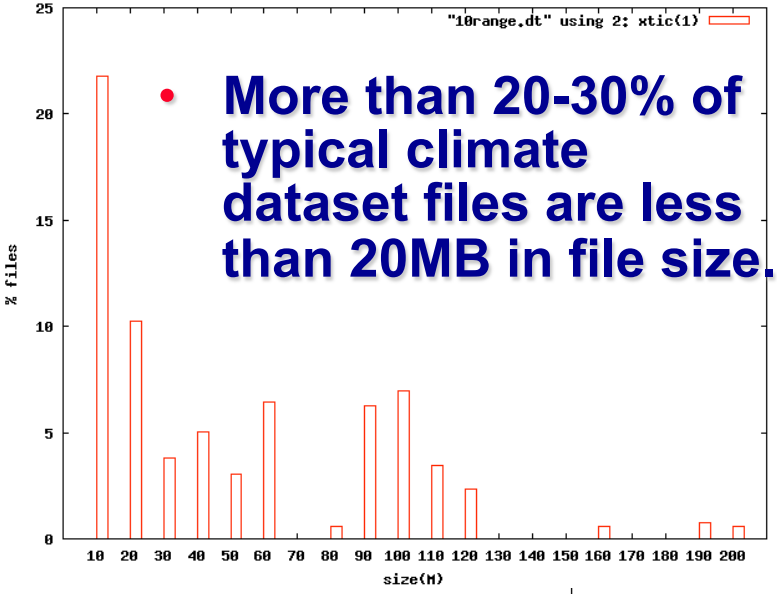
- Dynamic transfer parameter adjustment model
- Large-scale climate data analysis on Cloud computing with remote data access
- Climate analysis with remote data access over RDMA
- Climate dataset replication performance optimization
- Supercomputing 2011 – exploring climate data access over 100Gbps network

Dynamic transfer parameter adjustment model

Distribution of Climate Files – Characteristics of datasets

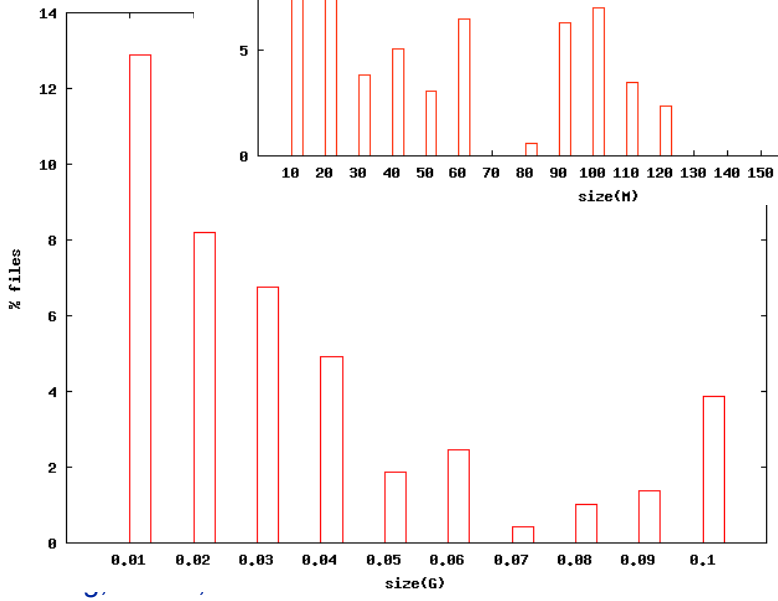


- More than 60-70% of typical climate dataset files are less than 200MB in file size.



- More than 20-30% of typical climate dataset files are less than 20MB in file size.

- Many files are still larger than 2GB.
- It's expected to be as large as 20+GB per file for CMIP-5 datasets.



End-to-end performance requirements for 100Gbps in climate dataset

- 100 Gbps = 12.5 GB/sec, 45TB in an hour, 1.08PB in a day
- 12.5 GB per second
  - 125 x 100MB files
  - 625 x 20MB files
  - 1,250 x 10MB files

- Extreme variance in file sizes affects end-to-end performance.
- Data i/o for so many files
  - Need an extensive data management
  - Need a big coordination between storage and network
    - Assuming 60MB/sec per spinning disk, 12.5GB/sec needs ~200 disks in full speed, opening many files at once

“Faster” Data Transfers

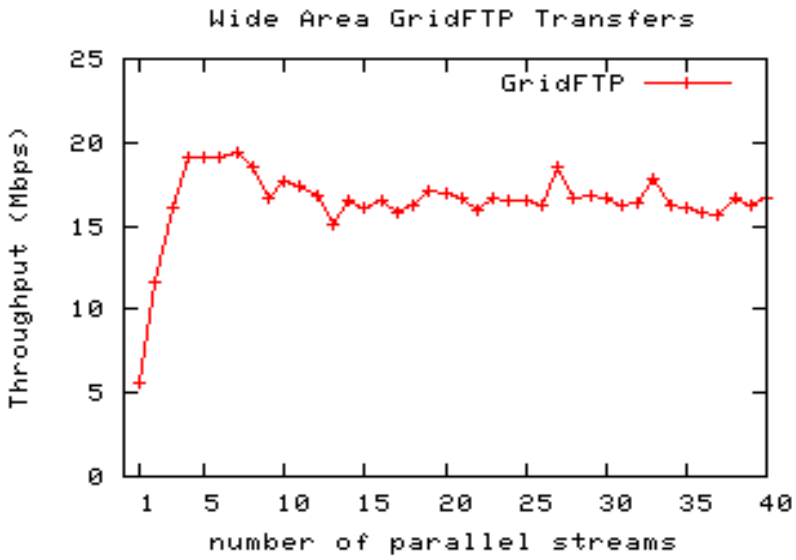
- Many solutions require kernel level changes
- Not preferred by the most domain scientists
- End-to-end bulk data movement
  - Latency issue

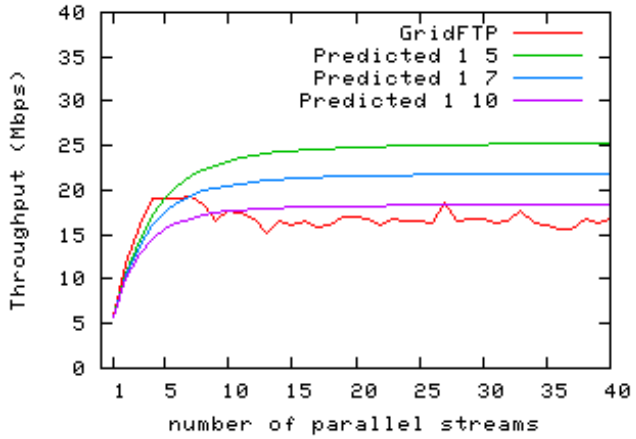
Application level tuning

- Application-level transfer protocol (i.e. GridFTP) and tune-up for better performance:
  - Using multiple streams
  - Adjusting buffer size
- Efficient utilization of available network capacity
- Level of parallelism in end-to-end data transfer
  - Number of parallel data streams
  - Number of concurrent data transfer operations
  - Multiple streams puts extra system overhead

Transfer parameter estimation

- Can we predict this behavior?
- We can come up with a good estimation for the parallelism level
  - Network statistics
  - Additional measurements
  - Historical data
- A model between RTT and the number of streams n





$$p'_n = p_n \frac{RTT_n^2}{c^2 MSS^2} = a'n^2 + b'$$

$$Th_n = \frac{n}{\sqrt{p'_n}} = \frac{n}{\sqrt{a'n^2 + b'}}$$

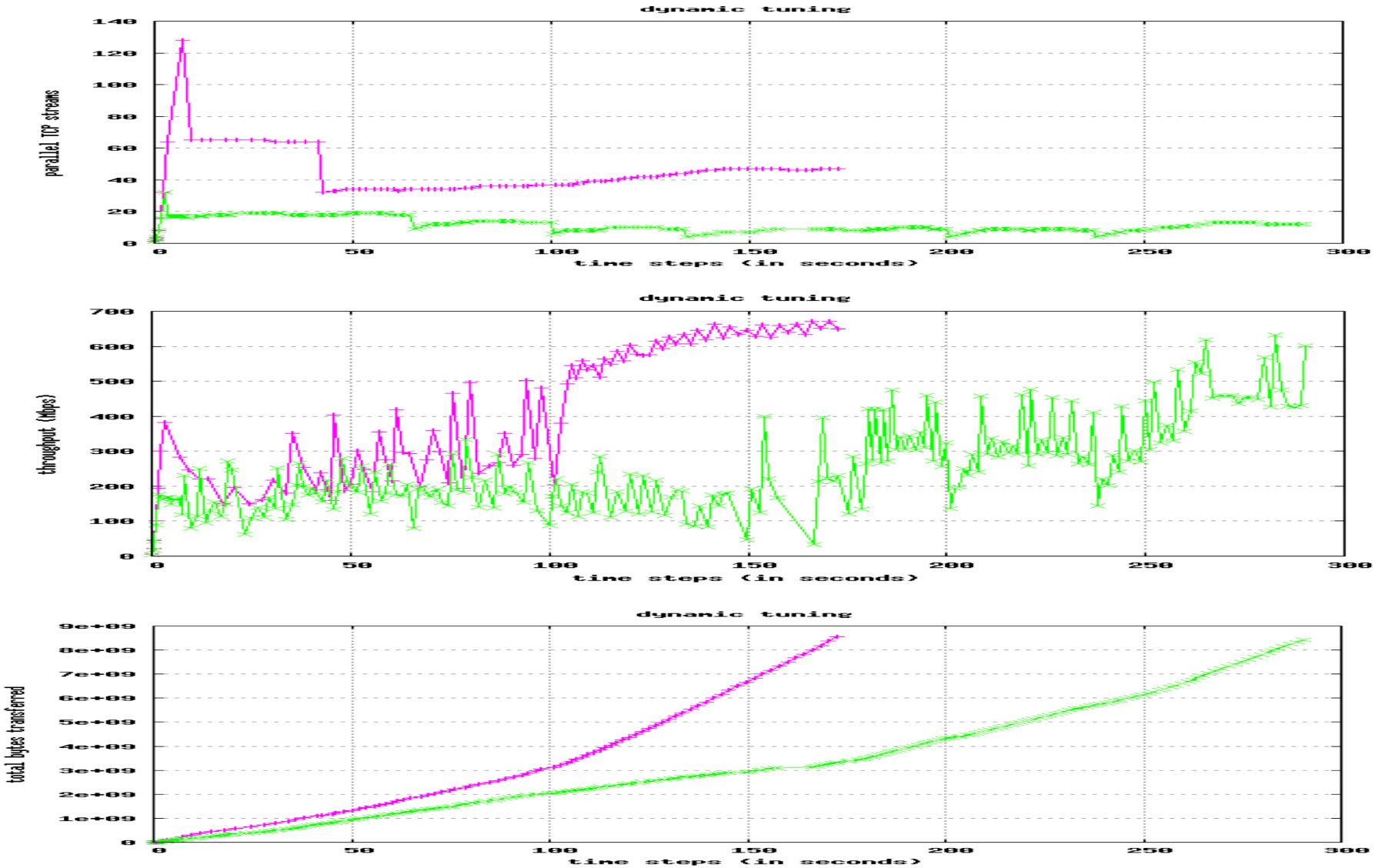
Transfer parameter estimation

- Might not reflect the current settings (dynamic environment)
  - What if network condition changes?
- Need multiple sample transfers (curve fitting)
  - Need to probe the system and make measurements with external profilers
- Does require a complex model for parameter optimization

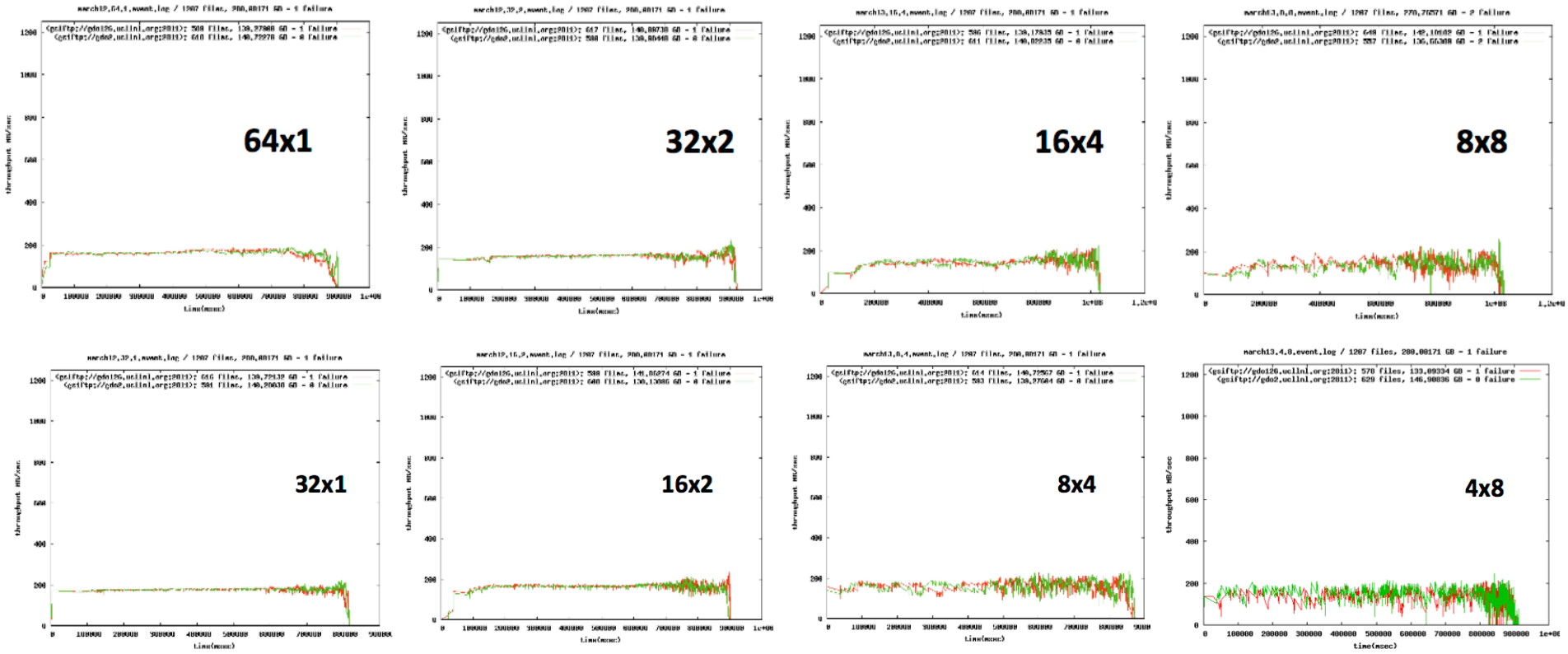
Adaptive tuning

- **Methods**
  - Instead of predictive sampling, use data from the actual transfers
  - Measure throughput for transferred data chunk
  - Gradually increase the number of streams until it comes to an equilibrium point
- **Advantages**
  - No need to probe the system and make measurements with external profilers
  - Does not require any complex model for parameter optimization
  - Adapts to changing environment
  - Fast start with exponentially increasing the number of streams
- **Disadvantages**
  - Overhead in changing parallelism level during the transfers

Test results on adaptive tuning

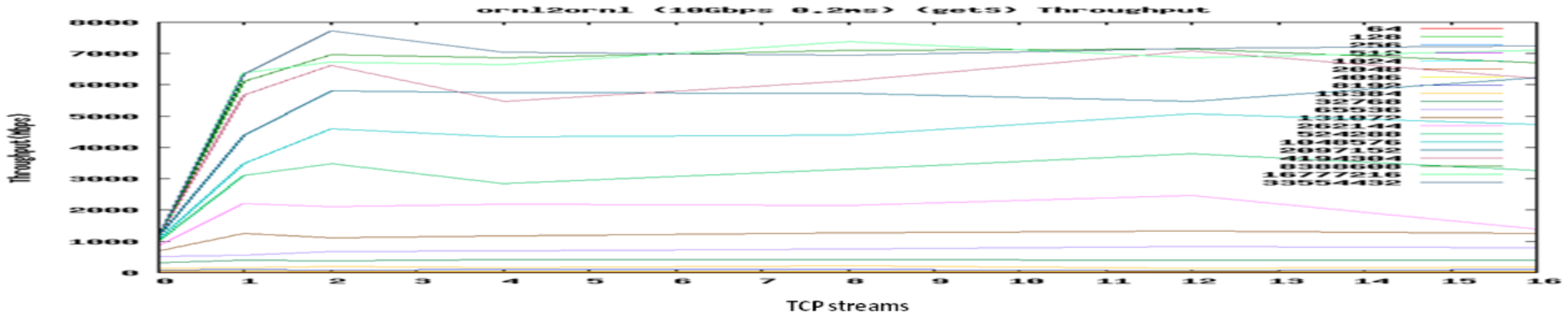


Observations (1)

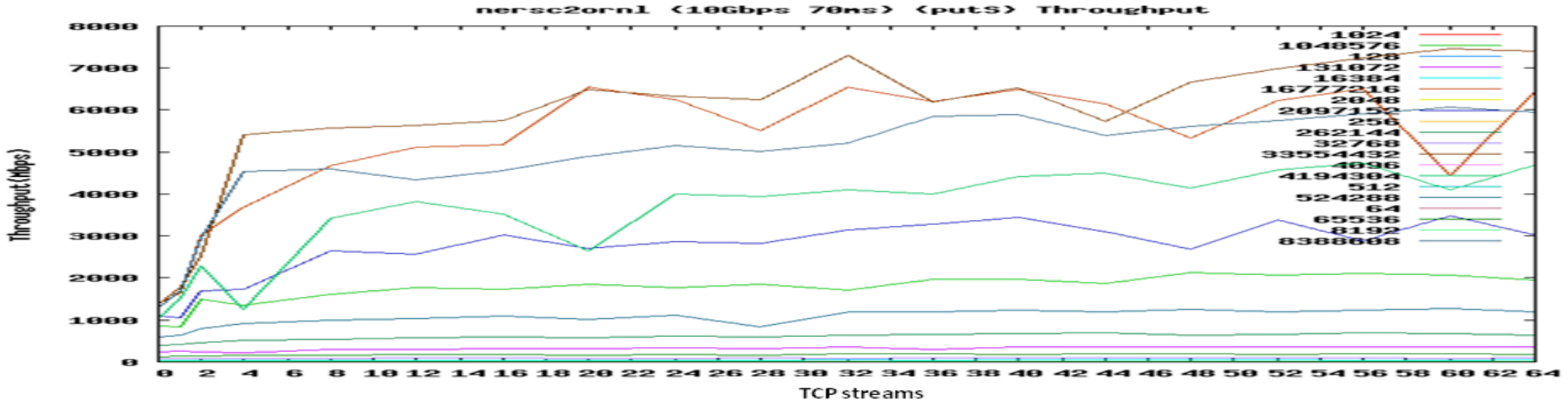


- Parallel streams vs. concurrent transfers
- Same number of total streams, but different number of concurrent connections

Observations (2)



(a)

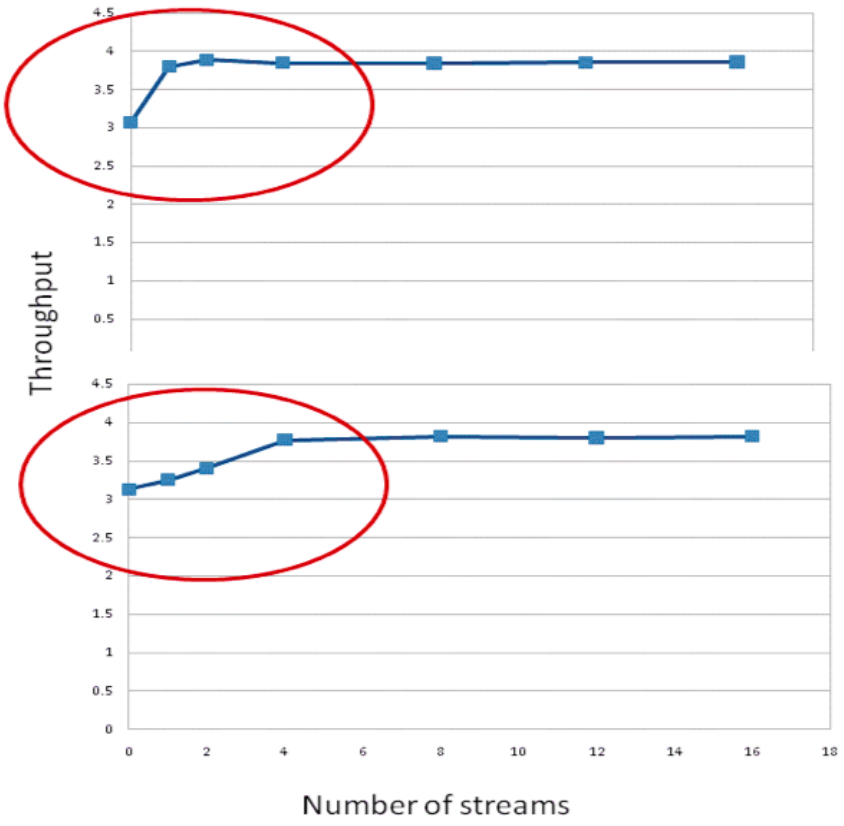


(b)

- **Latency directly affects the behavior of the throughput performance curve.**

Observations (3)

- The relationship between the number of multiple streams and the throughput gain can be approximated by a simple power-law model.
- Power-law approximation models the behavior of the multiple streams vs throughput in the first part where 80% of the achievable throughput is reached



Log-log graph: total throughput vs. number of streams

Simple throughput prediction model

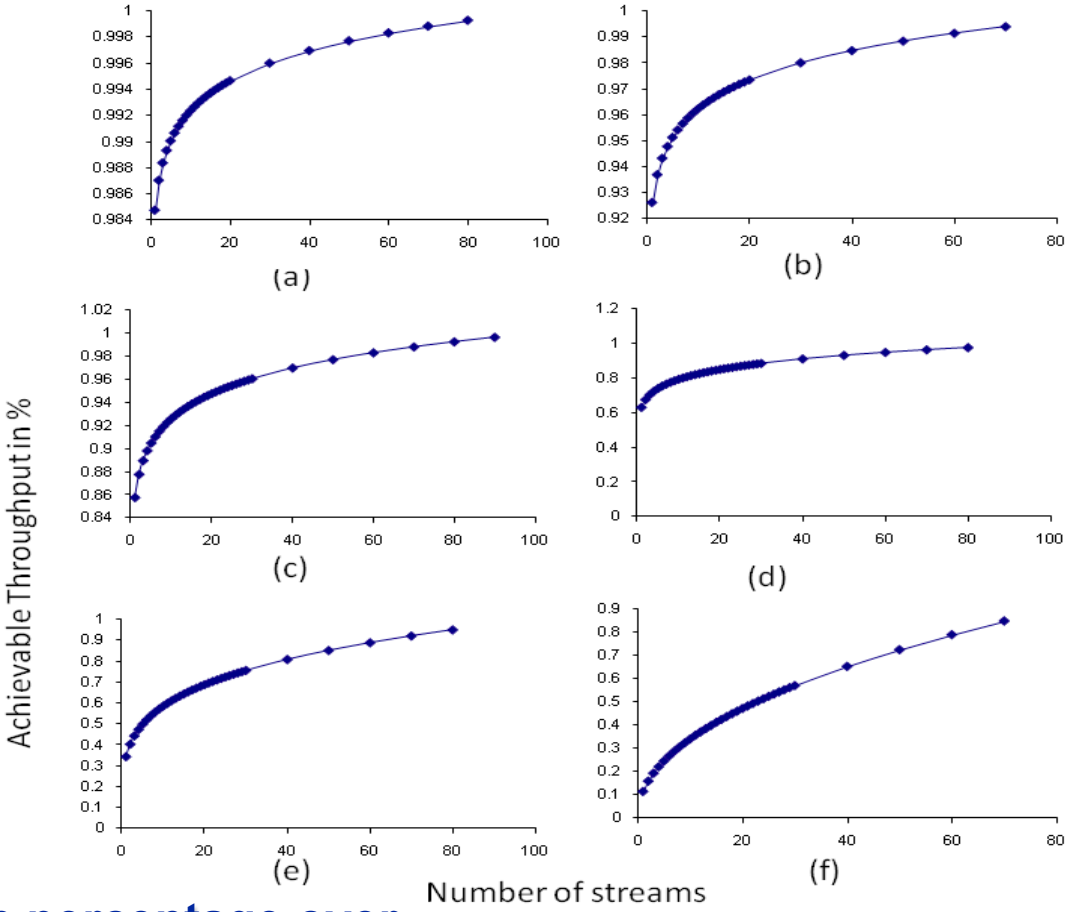
Power-Law model

$$T = (n / c)^{(RTT / k)}$$

80-20 rule-Pareto dist.

$$0.8 = (n / c)^{(RTT / k)}$$

$$n = (e^{(k * \ln 0.8 / RTT)}) \cdot c$$



Achievable throughput in percentage over the number of streams with low/medium/high RTT;
(a) RTT=1ms, (b) RTT=5ms, (c) RTT=10ms, (d) RTT=30ms, (e) RTT=70ms, (f) RTT=140ms (c=100, (n/c)<1, k =300 max RRT)

Supercomputing 2011

SC'11 demo

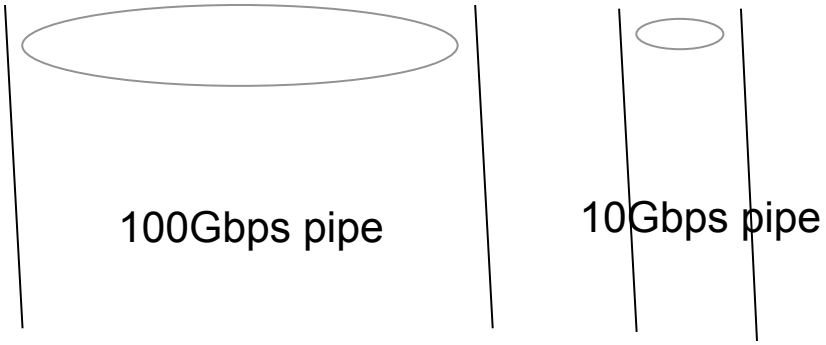
- Use of maximum available bandwidth for the movement of climate data over 100Gbps network
- Address data management challenges, in terms of
  - High bandwidth networks
  - Usability of existing transfer protocols and middleware tools
  - How applications can adapt and benefit from next generation networks
- Performance analysis of end-to-end data movement
  - Detailed profiling of the data transfer applications
  - Memory usage, number of context switches, time spent waiting I/O completion, user and system time, call graph of system calls, and time spent in each user operation
  - Expected that inefficient use of end system resources would be the major bottleneck in high-bandwidth networks

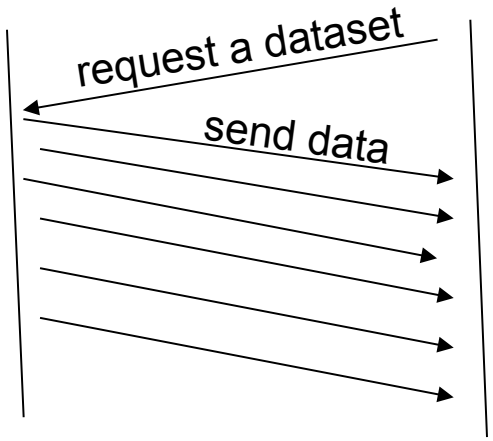
SC'11 demo

- **Expected challenges**
  - Irregular file size distribution in each dataset
    - Protocol overhead
  - Using existing tools in 100 Gbps networks
  - Performance problem and scalability issues
    - Management and tuning of multiple hosts
    - Multiple streams for increased utilization
    - Performance monitoring in host systems
    - Memory overhead / CPU usage
    - System bottleneck in end-to-end transfers
- **Measured performance of end-to-end data movement**
  - Developed a measurement tool for profiling and measuring end-to-end performance

High bandwidth networks

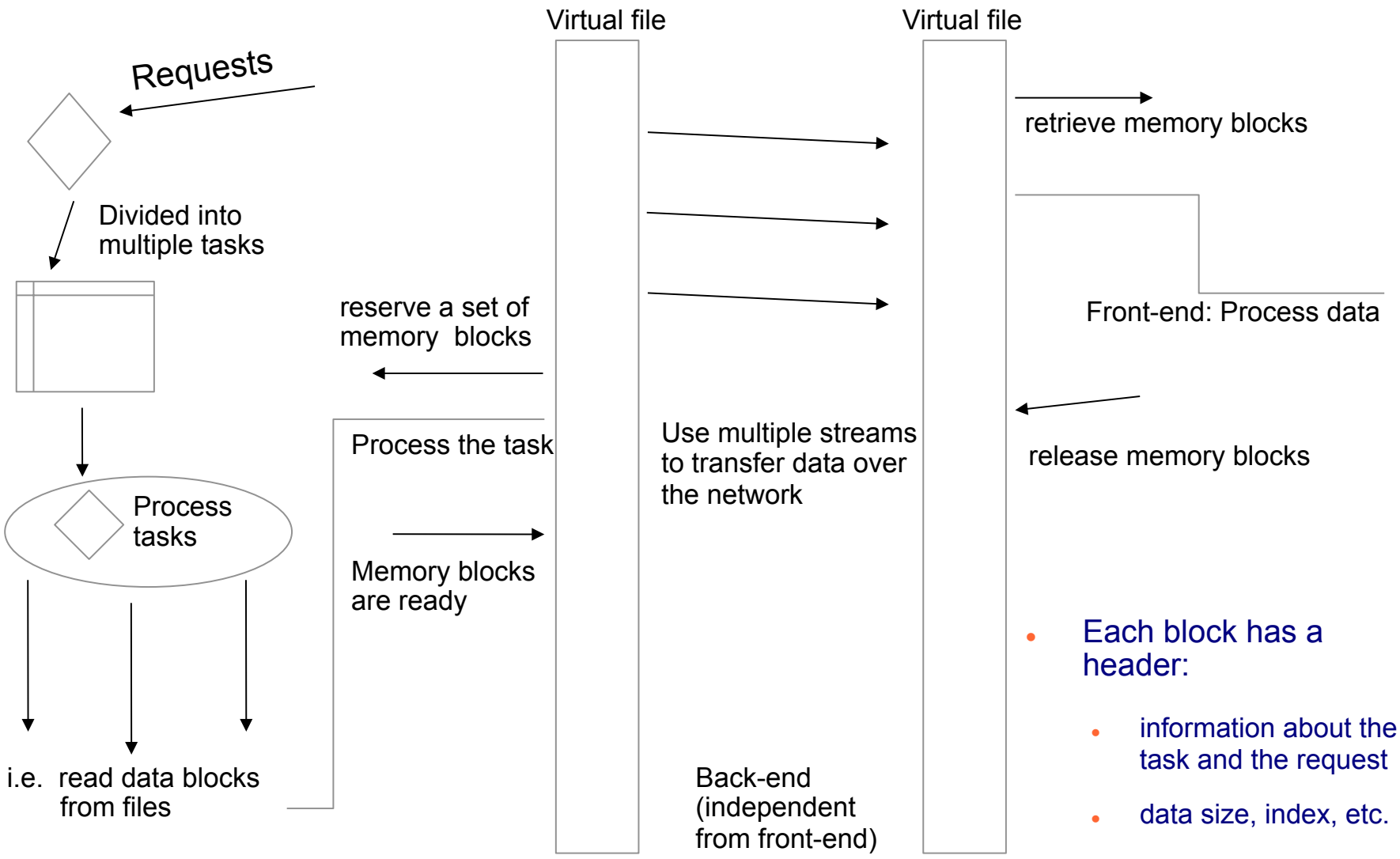
- **Same distance, and latency is still the problem**





- **Keep the pipe full?**
  - **Parallelism**
  - **Pipelining**
  - **???**
- **performance measurement**
  - **Minimize system overhead for network performance**
  - **Minimize the control messages**
  - **Aggregate requests**
  - **Pre-processing and post-processing at the end nodes**

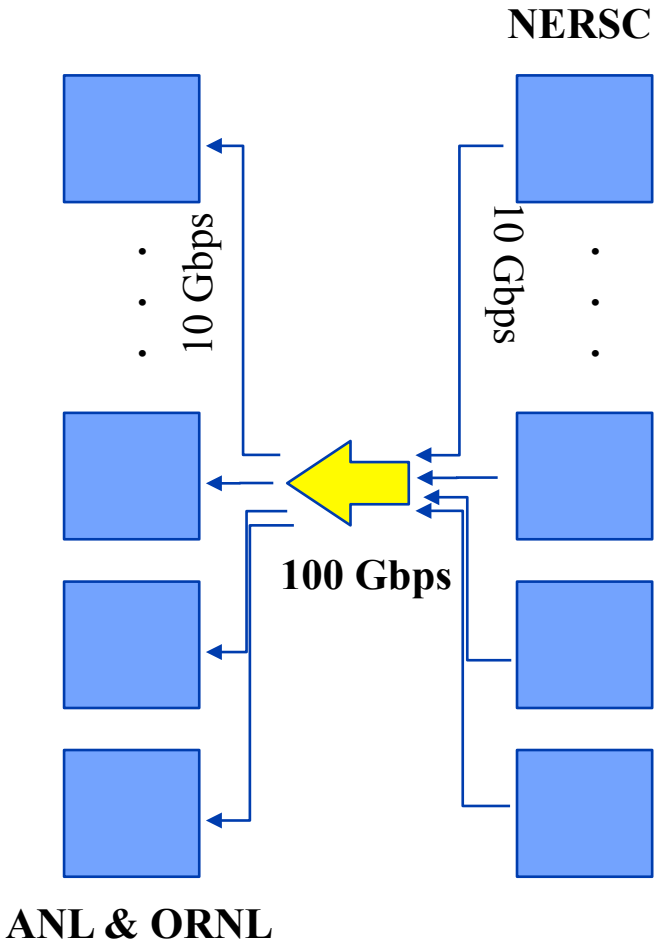
Data block communication



- Each block has a header:
  - information about the task and the request
  - data size, index, etc.

100Gbps Testbed for SC'11

- **Testbed for demo**
  - LBNL/NERSC
  - ANL/ALCF
  - ORNL/OLCF
  - Each node with 10 Gbps connection
- **IPCC AR4 CMIP3: ~35TB**
  - From NERSC to ANL over 100Gbps
    - Disk to memory
  - From NERSC to ORNL over 100Gbps
    - Disk to disk
  - Over TCP
  - No TCP tuning (rely on system param)
  - 4MB block size
  - 8 streams for each connection



Demo results

Scaling the Earth System Grid to 100Gbps Networks (LBNL)

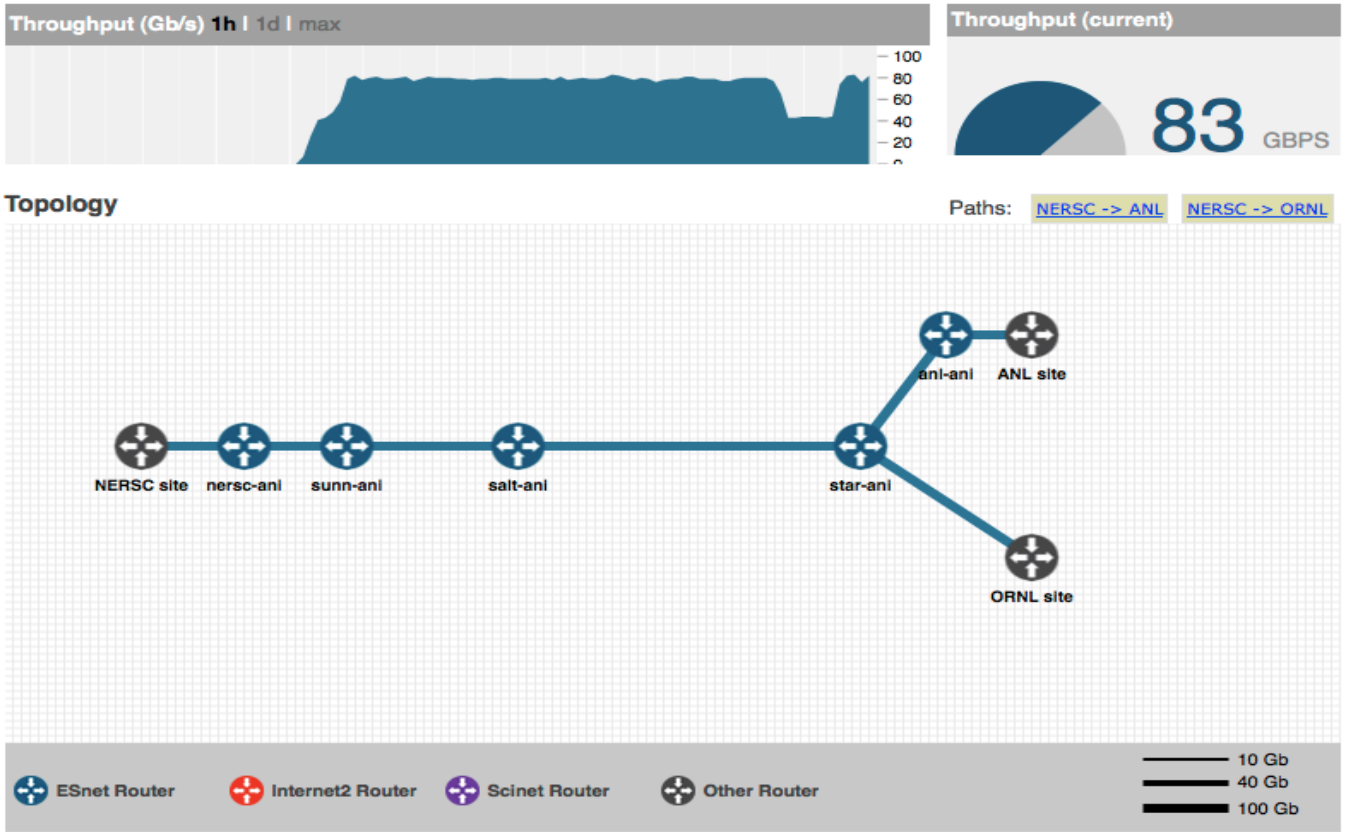
Presenters: Alex Sim and Mehmet Balman (LBNL Computational Research Division)

Venue: Booth 512

Climate change research is one of the critical data intensive sciences, and the amount of data is continuously growing. Climate simulation data is geographically distributed over the world, and it needs to be accessed from many sources for fast and efficient analysis and intercomparison of simulations. This demonstration will leverage a 100 Gbps link connecting the National Energy Research Scientific

more

Live demo stats



Current status